Semi-Supervised Video Desnowing Network via Temporal Decoupling Experts and Distribution-Driven Contrastive Regularization

Hongtao Wu[1], Yijun Yang[1], Angelica I. Aviles-Rivero[3], Jingjing Ren[1], Sixiang Chen[1], Haoyu Chen[1], and Lei Zhu[1,2]✉

[1] The Hong Kong University of Science and Technology (Guangzhou), China
[2] The Hong Kong University of Science and Technology, Hong Kong SAR, China
leizhu@ust.hk
[3] University of Cambridge, UK

Abstract. Snow degradations present formidable challenges to the advancement of computer vision tasks by the undesirable corruption in outdoor scenarios. While current deep learning-based desnowing approaches achieve success on synthetic benchmark datasets, they struggle to restore out-of-distribution real-world snowy videos due to the deficiency of paired real-world training data. To address this bottleneck, we devise a new paradigm for video desnowing in a semi-supervised spirit to involve unlabeled real data for the generalizable snow removal. Specifically, we construct a real-world dataset with 85 snowy videos, and then present a Semi-supervised Video Desnowing Network (SemiVDN) equipped by a novel Distribution-driven Contrastive Regularization. The elaborated contrastive regularization mitigates the distribution gap between the synthetic and real data, and consequently maintains the desired snow-invariant background details. Furthermore, based on the atmospheric scattering model, we introduce a Prior-guided Temporal Decoupling Experts module to decompose the physical components that make up a snowy video in a frame-correlated manner. We evaluate our SemiVDN on benchmark datasets and the collected real snowy data. The experimental results demonstrate the superiority of our approach against state-of-the-art image- and video-level desnowing methods. Our code and the dataset are available at https://github.com/TonyHongtaoWu/SemiVDN.

Keywords: Video desnowing · Semi-supervised learning · Mixture of experts · Contrastive learning

1 Introduction

Snow, one kind of adverse weather, frequently appears in outdoor videos. The degradation effects caused by snow particles and streaks severely impair the visibility of video frames and subsequently hinder the advanced performance of video processing algorithms in autonomous systems. Consequently, as an ill-posed inverse problem, a large quantity of desnowing methods are designed to

Fig. 1: The distribution shift of the synthetic snow and real snow.

Fig. 2: Left sub-figure: The proposed semi-supervised method trained using synthetic and real videos yields favorable results on snowy video samples captured in various scenarios, including forests, country roads and movies. **Right sub-figure:** Trade-off between PSNR performance v.s Runtime and GFLOPs on RVSD dataset [8].

prevent the perturbation of snow not only in single-image but also in videos. Early image snow removal methods tend to perform snow removal based on physical priors [4, 42, 45, 82]. Subsequently, deep neural networks like Convolutional Neural Networks (CNNs) and Transformers [10, 13, 14, 33, 36, 63, 69, 80] are introduced to remove the snow more sophisticatedly. According to the atmospheric scattering model [39], Chen *et al*. [8] constructed a video desnowing benchmark and the degradation caused by the snow could be formulated by:

$$\boldsymbol{I}_{snow}\left(\boldsymbol{x}\right)=\boldsymbol{J}\left(\boldsymbol{x}\right)\boldsymbol{T}\left(\boldsymbol{x}\right)+\boldsymbol{A}\left(\boldsymbol{x}\right)\left(1-\boldsymbol{T}\left(\boldsymbol{x}\right)\right)+\boldsymbol{S}\left(\boldsymbol{x}\right)\,, \tag{1}$$

where $\boldsymbol{I}_{snow}$ denotes the video deteriorated by snow, $\boldsymbol{J}$ is the corresponding clean video, $\boldsymbol{T}$ is the transmission map, $\boldsymbol{A}$ is the atmospheric light and $\boldsymbol{S}$ is the snow map. They also proposed the first network SVDNet to leverage temporal redundancy for snow removal task. Though such method has achieved success on synthetic benchmarks, better results in recovering real-world snowy videos are highly desired for deployment in real applications.

Unfortunately, as shown in Figure 1, due to the distribution shift between synthetic data and real-world data, it inevitably happens that the existing desnowing methods are impeded by unrealistic training data and fail to handle the real snow with unpredictable shapes and motions. More importantly, it's impractical to train these models with plenty of paired real-world data because variable weather conditions, object and camera position make it extremely complicated to align the videos from the real scene. To address the aforementioned issues, it's a natural practice to consider unlabeled real-world data into the training stage in a semi-supervised fashion.

In this work, we collect 85 unpaired real-world snowy videos for the training of the proposed model. We use the mixed set composed of synthetic and real data under the Mean-Teacher architecture to enhance its generalizable capability

across various real scenarios. Specifically, we introduce a Distribution-driven Contrastive Regularization to prevent the deep model from the perturbation of diverse snow shape and motion from synthetic and real data. To obtain ultra-positive samples, we utilize the GMM likelihood to capture the synthetic snow most similar to the real counterpart by approximating the distribution of real snow components. We maintain and highlight the snow-invariant information by replacing the snow-specific counterpart in ultra-positive samples and contrarily replacing the background in negative samples.

Furthermore, though SVDNet [8] manipulates the physical prior of Eq. 1, they lack the beneficial guidance to present an explicit decoupling on each component. To this end, we improve the vanilla transformer block to a physics-based counterpart by introducing temporal decoupling experts. These experts explicitly attend to different compositions of the degradation based on the physical formula, which provides the decomposed feature for the subsequent recovery. We also introduce temporal decomposition router aggregating complementary information within videos to explore the correlations between consecutive frames.

Our contributions can be summarized as:

- To the best of our knowledge, we present the first semi-supervised video desnowing framework named SemiVDN, which explicitly explores the beneficial knowledge from unlabeled data to improve the generalization capability of the deep model.
- We introduce a Prior-guided Temporal Decoupling Experts module to explicitly decompose the physical components considering inter-frame coherence for better snow removal.
- We also design a Distribution-driven Contrastive Regularization to mitigate the appearance difference between the synthetic and real data, and consequently maintain the desired snow-invariant information.
- Extensive experimental results on both synthesized videos and real-world snowy videos demonstrate that our network significantly outperforms other state-of-the-art snow removal methods. More importantly, it surpasses previous methods in trade-off and performance substantially and has a better generalization ability to benefit real-world applications as shown in Figure 2.

2 Related Work

2.1 Snow Removal Methods

Prior to the advent of deep learning [17, 19, 44, 55, 67, 71, 74], snow removal techniques [4, 42, 45, 82] predominantly relied on physics-based priors to address the snow removal challenge. In recent years, deep-learning-based methods [9, 11, 13, 14, 33, 36, 66, 70, 73, 80] have achieved impressive results for snow removal. JSTASR [13] proposed a snow removal algorithm that can jointly classify snow particles and remove the snow with different transparency. HDCW-Net [14] utilized a hierarchical decomposition paradigm, incorporating dual-tree wavelet transform and wavelet loss. DDMSNet [80] exploited semantic and depth

priors for image snow removal. Li *et al*. [33] proposed an online multi-scale convolutional sparse coding model for online snow removal. Previous research has primarily focused on single-image snow removal techniques, neglecting the complexities of video sequences under snowfall conditions. SVDNet [8] presents a video desnowing network with a snow-aware temporal aggregation module by integrating optical flow and snow features to guide the detection and removal of remaining snow within the video sequence. However, the aforementioned methods were only trained on synthetic data and may degenerate when deployed on real images caused by the distribution shift.

2.2 Semi-Supervised Learning

In recent years, semi-supervised learning [54, 56, 57, 83] has played an increasingly important role in tackling computer vision problems. Researchers introduced semi-supervised methods to learn real data patterns in image restoration tasks. Wei *et al*. [64] developed a semi-supervised image deraining model using a likelihood term from a parameterized distribution designed for residuals. S2VD [77] proposed a semi-supervised video deraining model with a dynamical rain generator. Recently, many semi-supervised methods have been developed, such as Mean-Teacher [52] and MixMatch [3]. Among them, the Mean-Teacher method often manipulates consistency regularization based on the high-quality pseudo-labels obtained by an exponential moving average network, which triggers its applications to vision tasks such as semantic segmentation [16] and image restoration [35, 59]. DMT-Net [35] utilized a disentangled-consistency network ensuring consistency between coarse predictions and refinements of real data for image dahazing. Wang *et al*. [59] leveraged a student-teacher framework via knowledge transfer for image super-resolution. To the best of our knowledge, semi-supervised learning has not yet been explored in the video snow removal task.

2.3 Mixture of Experts

Motivated by various successful cases of the Mixture of Experts (MoE) [26] in recent advances natural language processing (NLP) tasks [20, 48, 49, 58], especially the large language model (LLM), sparse MoE have been popular in high-level vision tasks [1, 2, 18, 21, 43, 60, 76] due to scaling up module capacity without sacrificing computational cost. Specifically, MoE involves a set of expert networks and a gating network, where gating scores from the gating network adjust the expert networks' outputs. In the community of low-level vision, DRSformer [15] introduces a mixture of experts feature compensators to perform a collaborative refinement of data and content sparsity for image deraining. Rather than adopting a sparse and discrete router, all the weights in our Temporal Decoupling Experts are continuously considered, while the physics-driven formula implicitly trains the corresponding experts.

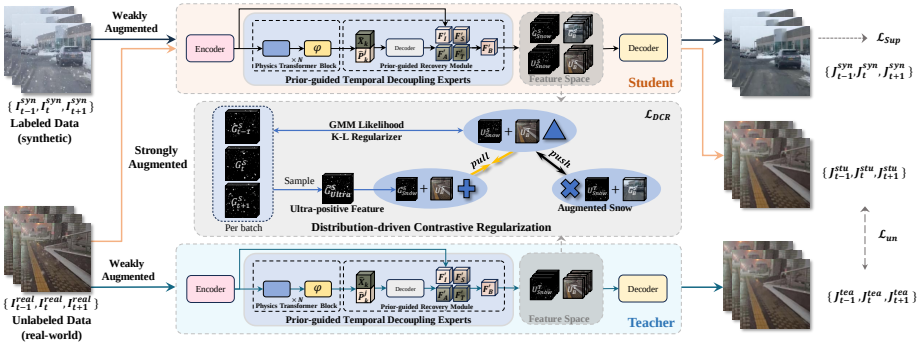

Fig. 3: The schematic illustration of our Semi-Supervised Video Desnowing Network (SemiVDN). SemiVDN is based on the mean teacher scheme with a student model and a teacher model. We first develop a Prior-guided Temporal Decoupling Experts (see Fig. 5) to decompose the physical components that make up a snow video in a temporal spirit. After that, we compute supervised losses for labeled data and unsupervised losses for unlabeled data. Based on the decomposed component features ($F_B^{'}$ and $F_S^{'}$) in representation space, we develop a Distribution-driven Contrastive Regularization to highlight the snow-invariant information by replacing the snow-specific feature in ultra-positive samples and replacing the background in negative samples.

2.4 Contrastive Learning

Contrastive learning, an efficient self-supervised learning method [12, 23, 40, 68, 72], aims to bring anchors closer to positive samples while distancing them from negative samples in the representation space. Some works have explored such the paradigm in low-level vision tasks [61, 65]. They adopted the original images as positive instances and the degraded images as negative instances, which were subsequently projected into the feature space via VGG [50] for contrastive learning. Semi-UIR [25] constructed a reliable bank to get the highest image quality samples as pseudo ground truth, which applied contrastive learning on unlabeled data. SVDNet [8] pioneered the application of contrastive learning in the desnowing task, based on the observation that distinct videos exhibit unique snow features, whereas identical videos maintain consistent snow features.

3 Methods

3.1 Network Architecture

Figure 3 illustrates the overall framework of the SemiVDN for video snow removal, which is constructed based on the Mean-Teacher fashion [52]. Specifically, we develop a video desnowing network (VDN) consisting of an encoder, a novel Prior-guided Temporal Decoupling Experts module, and a decoder. Given a snowy video sequence $\{\boldsymbol{I}_k \in \mathbb{R}^{3\times h\times w} \mid k \in [0, N_f)\}$, we adopt a universal backbone ConvNeXt [37] as the encoder to extract the feature maps of frames.

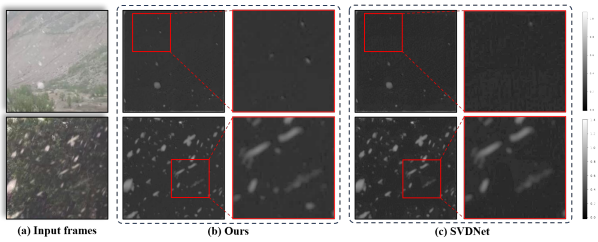

Fig. 4: Comparison of the snow layer decomposition results. It indicates our method can decouple more accurate and clean snow layers without background interference.

The extracted feature is further fed into our Prior-guided Temporal Decoupling Experts module, which aims to obtain the physical prior components that make up snow videos and remove undesired snow based on Eq. 1. After that, we feed the output desnowed background feature into a decoder to get the final prediction $\{\boldsymbol{J}_k \in \mathbb{R}^{3\times h\times w} \mid k \in [0, N_f)\}$. During the supervised stage, the labeled data is fed into the student network, and pixel-wise supervised loss is computed between the restored and clean frames. In the semi-supervised stage, we feed the unlabeled data into the student and teacher network, and compute the pixel-wise consistency loss, perceptual contrastive loss and prior losses to regularize the student network. Moreover, we also exploit Distribution-driven Contrastive Regularization Loss to prevent the model from the negative effects of the distribution gap between synthetic and real data. In the testing stage, we utilize the student network to predict the desnowed results from the input frames.

3.2 Prior-Guided Temporal Decoupling Experts

Previous works [8, 13] tended to remove snow by mimicking its physical model in image and video. For example, SVDNet [8] decouples the fused feature and derives several physical features by the convolutional layers based on the formula. As shown in Figure 4, this approach frequently fails to capture clean decoupled features (i.e. snow lay feature). To enhance the decoupling ability of the network, we first define a physics transformer block with a set of experts that decomposes the backbone feature into several representative physics-specific features.

Physics Transformer Block. Figure 5 illustrates the detailed procedure of the proposed Prior-guided Temporal Decoupling Experts module. The encoder obtains feature maps of frames $\{\boldsymbol{X}_k \in \mathbb{R}^{h/4\times w/4\times c} \mid k \in [0, N_f)\}$, and each $\boldsymbol{X}$ is individually performed overlapped patch embedding. Then all patches are linearly embedded into tokens $\boldsymbol{Y} \in \mathbb{R}^{(N_f\cdot m)\times d}$, where m is the number of tokens in one frame and d is the token channel. Then, $\boldsymbol{Y}$ is fed into transformer blocks for physics-dependent information separation. In the physics transformer blocks, we first improve the first transformer's feed-forward network by a fusion feed-forward network [34] to enhance feature fusion. Inspired by the popular design of sparse MoEs [20, 31, 43, 46], we incorporate our Temporal Decoupling Experts module into the second transformer of the Physics Transformer Block

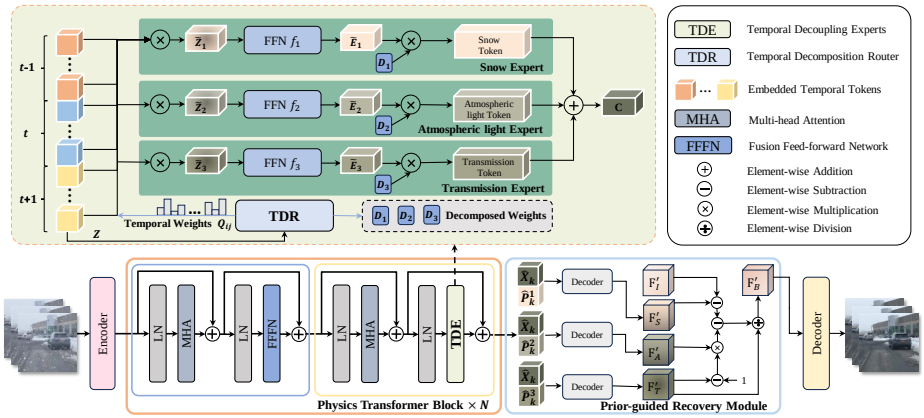

Fig. 5: Illustration of the proposed Prior-guided Temporal Decoupling Experts framework. Given an input snowy sequence, Physics Transformer Block (PTB) accepts encoded features as input and employs Temporal Decoupling Experts module to generate physics-specific components (i.e. S, A and T) for recovery. Specifically, we utilize the Temporal Decomposition Router to compute the temporal weights $\mathbf{Q}_{ij}$ from the temporal dimension, which are subsequently employed to compute a linear combination of all input temporal tokens and $\mathbf{Q}_{ij}$. Then each expert (an MLP in this work) processes its temporal adaptive tokens to obtain the corresponding output component tokens $\tilde{\mathbf{E}}_j$. Finally, we employ the decomposed weights from Temporal Decomposition Router to convexly combine all the component tokens. The output combined features $\hat{X}_k$ and physics-specific features $\hat{P}_k^j$ are subsequently input into the Prior-guided Recovery Module and the decoder to generate the ultimate desnowed results.

by replacing its MLP blocks. The proposed module consists of specific experts corresponding to diverse physics components, that are snow expert, transmission expert and atmospheric light expert. We denote the input tokens for one sequence by $\boldsymbol{Z} \in \mathbb{R}^{(N_f \cdot m) \times d}$. Every Temporal Decoupling Experts module utilizes a set of n expert functions, specifically denoted as $\left\{f_j : \mathbb{R}^d \to \mathbb{R}^d\right\}_{j=1}^{n}$. Each expert will process a temporal adaptive token, and each token has a corresponding d-dimensional vector of parameters, denoted as $\boldsymbol{\Gamma} \in \mathbb{R}^{d \times n}$. Then we can get the temporal adaptive weight $\mathbf{Q}_{ij}$ based on the temporal dimension $N_f \cdot m$ from the Temporal Decomposition Router as follows:

$$\mathbf{Q}_{ij} = \frac{\exp\left((\mathbf{Z}\boldsymbol{\Gamma})_{ij}\right)}{\sum_{i=1}^{N_f \cdot m} \exp\left((\mathbf{Z}\boldsymbol{\Gamma})_{ij}\right)}. \tag{2}$$

Consequently, the weighted tokens $\tilde{\mathbf{Z}} \in \mathbb{R}^{n \times d}$ are the convex combination of $N_f \cdot m$ input tokens and $\mathbf{Q}_{ij}$, which contains adaptive temporal information from the input N_f frames:

$$\tilde{\mathbf{Z}} = \mathbf{Q}^\top \mathbf{Z}. \tag{3}$$

Then, the corresponding expert function is applied to each temporal adaptive token $\tilde{\mathbf{Z}}_j$ to obtain the output component tokens:

$$\tilde{\mathbf{E}}_j = f_j\left(\tilde{\mathbf{Z}}_j\right) . \tag{4}$$

We can perform dynamic decoding of physics-specific features with the obtained representative component tokens and decomposed weights based on the Temporal Decomposition Router. Finally, the output tokens $\mathbf{C}$ are computed as a convex combination of all output component tokens:

$$\mathbf{D}_{ij} = \frac{\exp\left((\mathbf{Z}\boldsymbol{\Gamma})_{ij}\right)}{\sum_{j=1}^{n}\exp\left((\mathbf{Z}\boldsymbol{\Gamma})_{ij}\right)}, \mathbf{C} = \mathbf{D}\tilde{\mathbf{E}}, \tag{5}$$

where $\mathbf{D}$ is the decomposed weights, *i.e.*, the softmax results across the expert dimension of $\mathbf{Z}\cdot\boldsymbol{\Gamma}$. Sparse MoE algorithms typically have a discrete nature, making them non-differentiable. This can lead to missing information, such as token dropping and expert unbalance when using classical routing mechanisms. In contrast, our Temporal Decoupling Experts employ continuous and differentiable operations. They effectively leverage the information from all temporal tokens and experts, resulting in the extraction of physics-specific features. Then, the combined tokens sequence $\mathbf{C}$ and each component tokens are temporally transformed to obtain spatial feature maps $\{\hat{\boldsymbol{X}}_k \in \mathbb{R}^{h/4\times w/4\times c} \mid k \in [0, N_f)\}$ and $\{\hat{\boldsymbol{P}}_k^j \in \mathbb{R}^{h/4\times w/4\times c} \mid k \in [0, N_f), j \in [1, n]\}$, respectively. After that, each component feature is, respectively, concatenated with the combined counterpart and then fed into their specific decoder to obtain the enhanced component features. The three decoders are composed of three convolutional layers with upsampling, respectively. Finally, these enhanced features are utilized for the final recovery in the subsequent prior-guided recovery module.

Prior-Guided Recovery Module. We employ the Eq. 1 for the simultaneous removal of snow and haze in frames. This model facilitates the decomposition of frames into three distinct components S, A, and T in the feature space. According to the Eq. 1, the prior-guided recovery process can be formulated as:

$$F_B^{'} = \frac{F_I^{'} - F_S^{'} - (1 - F_T^{'})F_A^{'}}{F_T^{'} + \beta}, \tag{6}$$

where $F_I^{'} \in \mathbb{R}^{N_f\times c\times h\times w}$ is the encoded input feature, $F_S^{'} \in \mathbb{R}^{N_f\times c\times h\times w}$ is the snow feature, $F_T^{'} \in \mathbb{R}^{N_f\times c\times h\times w}$ is the transmission feature, $F_A^{'} \in \mathbb{R}^{N_f\times 1\times h\times w}$ is the global atmospheric light feature and the hyper-parameter β is set to 10^{-8}. Finally, we project the output desnowed feature $F_B^{'}$ into RGB space with a convolution layer to obtain the snow-free frame $\mathbf{J}$.

3.3 Semi-Supervised Video Snow Removal

In order to enhance the generalization ability of our model across real-world data, we introduce semi-supervised learning (SSL) in video snow removal tasks.

SSL enables a learning system to explore complementary information from both labeled synthesized and unlabeled real-world data. As illustrated in Figure 3, our SSL framework follows the typical setup [52]. In the training process, the student network is updated by minimizing the supervised losses and unsupervised losses, while the teacher network is updated by the exponential moving average (EMA):

$$\theta^{'}_{teacher} = \eta\theta_{teacher} + (1-\eta)\theta_{student}\,, \tag{7}$$

where the EMA decay η is empirically set as 0.99. With the adoption of this update strategy, the teacher model can promptly aggregate weights that have been acquired in prior training steps.

Supervised Loss. To constrain the outputs of the student network, we adopt the Charbonnier loss [7] and the perceptual loss [29] to improve the visual quality of the restored results. While the L1 Charbonnier loss is commonly utilized, the perceptual loss is to quantify the disparity between the features of the prediction and the ground truth. We extract features from the 3-rd, 8-th and 15-th layers of the pretrained VGG-16 [50] to calculate the perceptual loss. We also introduced the Focal Frequency Loss [27] to focus the model on the response of different regions in the frequency spectrum to varying artifacts in the image. The overall supervised loss is formulated as:

$$\mathcal{L}_{\text{sup}} = \mathcal{L}_{pixel} + \lambda_1\mathcal{L}_{perceptual} + \lambda_2\mathcal{L}_{Frequency}\,, \tag{8}$$

where λ_1 and λ_2 are the balancing hyper-parameters, empirically set as 0.03 and 10, respectively.

Unsupervised Loss. Firstly, we adopt the pixel level Charbonnier loss $\mathcal{L}^{'}_{pixel}$ as the unsupervised teacher-student consistency loss to ensure that the two networks generate consistent results. Secondly, we follow [61,65] to incorporate the perceptual contrastive loss by constructing the corresponding perceptual features of J^{stu}_t and J^{tea}_t, and $\hat{I}^{real}_t$ as the anchor, positive and negative samples, respectively. Furthermore, inspired by [13,14,32,47], we employ the dark channel prior (DCP) [24] loss $\mathcal{L}_{DCP}$ and the total variation loss $\mathcal{L}_{TV}$ as the prior losses, which regularize student network to produce results J^{stu}_t with similar statistical characteristics of the clear images. The overall unsupervised loss is expressed as:

$$\mathcal{L}_{un} = \lambda_3\mathcal{L}^{'}_{pixel}(J^{stu}_t, J^{tea}_t) + \lambda_4\mathcal{L}_{cl}(\hat{I}^{real}_t, J^{tea}_t, J^{stu}_t) + \lambda_5\mathcal{L}_{DCP} + \lambda_6\mathcal{L}_{TV}\,, \tag{9}$$

where $\hat{I}^{real}_t$ denotes the strongly augmented unlabeled degraded video sequence, J^{stu}_t and J^{tea}_t denote the snow-free result predicted by student model and teacher model, respectively. While λ_3, λ_4 ,λ_5 and λ_6 are the balancing hyper-parameters, empirically set as 2, 0.1, 0.1 and 0.5. To get the description of unsupervised loss functions, please refer to the *Supplementary File* in detail.

Finally, the overall optimization objective of the student network can be formulated as minimizing the following loss:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{sup}} + \mu\mathcal{L}_{un}\,. \tag{10}$$

Inspired by [16,35], we apply a time-dependent Gaussian warming up function to update the weight $\mu : \mu(r) = \mu_{\max} e^{\left(-5(1-r/r_{\max})^2\right)}$, where r denotes the current training iteration and $r_{\max}$ is the maximum training iteration.

3.4 Distribution-Driven Contrastive Regularization

Consistency regularization struggles to mitigate the distribution gap between synthesized data and real data [25]. Such the domain gap between real and synthetic snow unexpectedly results in numerous incorrect pseudo-labels. The false information will be accumulated by overfitting on such pseudo-labels. In order to tackle the aforementioned concern, we introduce contrast learning to shift the attention of the network to the recovery of the snow-invariant background details of the snowy video.

As shown in Figure 3, base on our physics transformer block, we can get the pair of the background feature G_B^S and snow feature G_{Snow}^S from labelled data and the background feature U_B^S and snow feature U_{Snow}^S from unlabelled data in student network. Additionally, the background feature U_B^T and snow feature U_{Snow}^T are obtained from unlabeled data in the teacher network. According to the synthesis formula of snow $\boldsymbol{I}_S(\boldsymbol{x}) = \boldsymbol{J}(\boldsymbol{x}) + \boldsymbol{S}(\boldsymbol{x})$, we recombine the background features and snow layer features from labelled and unlabelled data. In order to maintain and highlight the snow-invariant information, we replace the snow-specific counterpart in positive samples and contrarily replace the background in negative samples. Specifically, we set the U_B^S and G_{Snow}^S as positive samples, U_B^T and U_{Snow}^S as anchor samples, G_B^S and augmented U_{Snow}^T as negative samples.

Due to the distribution differences between synthesized and real snow, the snow layers generated by our network often exhibit noticeable differences. Therefore, our objective is to acquire an ultra-positive sample, representing the synthetic snow layer that closely resembles the characteristics of real snow layers, to serve as a positive sample. Since the real snow generally contains inherent varied structures due to their different generation states and observed perspectives, they can be represented by a Gaussian Mixture Model (GMM). Employing GMM enables precise approximation of the distribution of real snow layers, effectively capturing the diverse modes within the data. The distribution of snow layers can be represented as:

$$\nu \sim \sum_{p=1}^{K} \pi_p \cdot \mathcal{N}(\nu \mid \alpha_p, \Sigma_p) \,, \tag{11}$$

where π_p, α_p, Σ_p denote the mixture coefficients, Gaussian distribution means and variances, respectively. By leveraging the GMM, we can effectively quantify the distribution of real and synthetic snow layers. This enables us to compute the Kullback-Leibler (KL) Divergence, which serves as a measure of dissimilarity between the Gaussian mixture module obtained from the real snow layer and diverse synthetic snow layers. Through a selection process that prioritizes the minimum KL divergence, we are able to identify the ultra-positive synthesized

Fig. 6: Samples of the proposed real-world video dataset for video snow removal.

snow layers $\hat{G}^S_{Ultra}$ that closely mirrored the distribution characteristics found in real snow layers U^S_{Snow}. After constructing the positive and negative samples, we can calculate the distribution-driven contrastive loss as follows:

$$\mathcal{L}_{DCR} = \frac{\mathcal{L}_{\mathrm{L}_1}\left(U^T_B + U^S_{Snow}, U^S_B + \hat{G}^S_{Ultra}\right)}{\mathcal{L}_{\mathrm{L}_1}\left(U^T_B + U^S_{Snow}, G^S_B + Aug\left(U^T_{Snow}\right)\right) + \varepsilon}, \tag{12}$$

where the hyper-parameter ε is set to 10^{-7}, $\mathcal{L}_{\mathrm{L}_1}(x, y)$ is the ℓ_1-distance loss between x and y, and the weight of this loss is set to 0.1. Eventually, we incorporate the $\mathcal{L}_{DCR}$ into the L_{un} to derive $L^{'}_{un}$.

4 Experiments

4.1 Real-World Snow Video Datasets

In order to tackle the aforementioned issue of lacking suitable video datasets to generalize the performance of desnowing algorithms in real-world snow video, we create the first video snow removal dataset Realsnow85, which is incorporated into the training processing of the semi-supervised video desnowing network. This dataset serves as a resource for researchers in this field to develop and test novel methods for the removal of snow from video data. To collect videos for training and testing, we select the snowy video data from the Internet. As shown in Figure 6, we capture different video backgrounds, such as cities, parks, villages and nature. In order to enable our model to cope with various snowfall and lighting conditions, we also considered different snowfall levels and lighting scenarios. In addition, we conduct a comprehensive experiment to evaluate our desnowing network on the Realsnow85 dataset, encompassing 85 videos that exhibit diverse scenes, resolutions, and degradation issues. Among these videos, 60 videos are utilized for training the network, while the other 25 videos are employed for testing and evaluating. Following [9, 10], we use Neural Image Assessment(NIMA) [51] and Multi-scale Image Quality Transformer(MUSIQ) [30] as the Non-reference Image Quality Assessment metrics to quantitatively compare the performance of real-world snow degraded video restoration.

Table 1: Quantitative comparisons between our network and other methods on synthetic datasets and real-world datasets. Bolded and underlined values indicate the best and the second-best performance, respectively.

Method	Type	Venue	Synthetic Datasets			Real-world Datasets	
			PSNR↑	SSIM↑	LPIPS ↓	NIMA↑	MUSIQ ↑
Input	-	-	18.37	0.7792	0.3095	4.075	48.64
JSTASR [13]	Image	ECCV 2020	22.08	0.8280	0.2336	4.173	48.82
HDCW-Net [14]	Image	ICCV 2021	22.63	0.8592	0.2010	4.208	47.54
Snowformer [10]	Image	arXiv 2022	24.01	0.8939	0.1219	4.215	49.78
SVDNet [8]	Video	ICCV2023	<u>25.06</u>	<u>0.9210</u>	<u>0.0842</u>	4.220	<u>50.78</u>
TransWeather [53]	Image	CVPR2022	23.11	0.8543	0.2086	4.182	48.06
WeatherDiffusion [41]	Image	TPAMI 2023	22.01	0.8621	0.1539	4.106	48.87
ViWS-Net [70]	Video	ICCV2023	24.43	0.8922	0.1142	<u>4.238</u>	50.56
MPRNet [79]	Image	CVPR2021	24.27	0.8960	0.1266	4.206	50.08
Restormer [78]	Image	CVPR2022	24.34	0.8929	0.1164	4.218	50.34
IR-SDE [38]	Image	ICML2023	22.71	0.8749	0.1168	4.099	47.45
IconVSR [5]	Video	CVPR2021	22.35	0.8482	0.2034	4.185	49.27
BasicVSR++ [6]	Video	CVPR2022	22.64	0.8618	0.1868	4.221	49.97
AECR-Net [65]	Image	CVPR2021	22.95	0.8530	0.1925	4.188	49.81
JRGR [75]	Image	ICCV2021	23.73	0.8729	0.1427	4.139	48.63
S2VD [77]	Video	CVPR2021	24.02	0.8761	0.1513	4.156	49.61
SemiVDN	Video	-	**25.68**	**0.9254**	**0.0785**	**4.259**	**51.57**

4.2 Implementation Details

Our network is trained on NVIDIA RTX 4090 GPUs and implemented on the Pytorch platform. The number of frames per video clip is three. Each input frame is randomly cropped to a spatial resolution of 256×256. The total number of the training iteration is 300K. We use the AdamW optimizer and the polynomial scheduler. The initial learning rate of our main network is set to 1×10^{-4} with a batch size of 4. We set the number of GMM components to be three.

4.3 Comparison with State-of-the-Art Methods

Compared Methods. To evaluate the effectiveness of the proposed method, we compare it against 15 state-of-the-art methods, including four snow removal methods [8,10,13,14], three adverse weather restoration methods [41,53,70], five fully-supervised restoration methods [5, 6, 38, 78, 79], and three semi-supervised restoration methods [65, 75, 77]. For a fair comparison, we implemented these fully-supervised methods by their official codes following [8] and retrained them on the RVSD dataset. For all compared semi-supervised methods, we follow the same setting of our method to retrain them on a training set, which contains the training set of the RVSD dataset and the training set of our proposed real dataset. Follow [22,28], we employed the peak signal-to-noise ratio (PSNR), the structural similarity index (SSIM) [62], and the learned perceptual image patch similarity (LPIPS) [81] to quantitatively compare the performance between different methods. The average scores of the three metrics are computed for all the frames between the predicted results and the ground truths in the testing set.

Synthetic Datasets. Table 1 reports the quantitative results of our network and 15 state-of-the-art methods on the RVSD test dataset. Among these methods, SVDNet stands out as the most competitive with the highest PSNR score of 25.06 dB, the highest SSIM score of 0.9210, and the lowest LPIPS score of 0.0842. Instead, our network outperforms SVDNet, evidenced by the higher PSNR and SSIM scores of 25.68 dB and 0.9254 respectively, as well as a lower LPIPS score

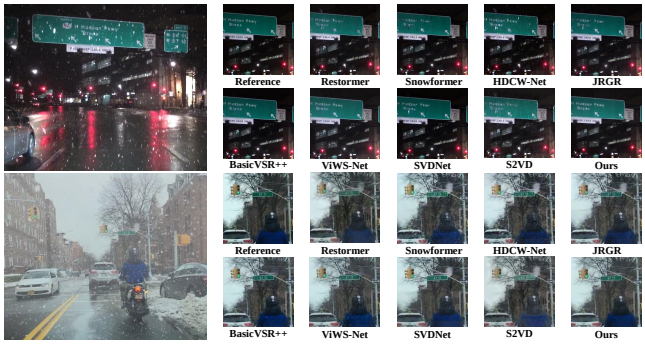

Fig. 7: Visual comparisons of desnowed results produced by our network and state-of-the-art desnowing methods for input video frames from the RVSD dataset.

Fig. 8: Visual comparisons of desnowed results produced by our network and state-of-the-art video desnowing methods for input video frames from real-world snowy videos.

of 0.0785. These results highlight the exceptional performance of our network in effectively removing snow and preserving image quality.

Real-World Datasets. As shown in the comprehensive results presented in the Table 1, our network also outperforms alternative methods in non-reference image quality evaluation metrics such as NIMA and MUSIQ. The comparison clearly demonstrates that our network excels in restoring images with superior quality, offering clearer content and enhanced perceptual fidelity when compared to other methods in real snowy scenarios.

Qualitative Comparison. Fig. 7 visually compares snow removal results predicted by our network and state-of-the-art methods from the RVSD dataset. Compared with other approaches, our network demonstrates superior performance in restoring the original background images by effectively eliminating snow and haze from input video frames. To further validate its efficacy on real-world data, we conduct a comparative analysis of different methods on snowy videos from our Realsnow85 testing set, as illustrated in Figs. 8. The results clearly indicate that our network excels in removing real snow and haze, while also successfully recovering obscured background details. Conversely, other methods tend to retain certain levels of snow and haze in their desnowed outputs.

4.4 Ablation Study

Baseline Design. To analyze the effectiveness of our SemiVDN, we conduct ablation studies to reveal the influence of three key components in our method,

Table 2: Quantitative results of our network and constructed baseline networks ("M1" to "M3") of the ablation study on synthetic datasets and real-world datasets.

Method	TDE	SST	DCR	PSNR↑	SSIM↑	LPIPS ↓	NIMA ↑	MUSIQ↑
M1				24.41	0.9116	0.0932	4.165	49.53
M2	✓			25.16	0.9217	0.0822	4.212	50.69
M3	✓	✓		25.29	0.9237	0.0806	4.239	51.05
SemiVDN	✓	✓	✓	**25.68**	**0.9254**	**0.0785**	**4.259**	**51.57**

i.e., the temporal decoupling experts (TDE), the semi-supervised training (SST), and the Distribution-driven Contrastive Regularization (DCR) of our SemiVDN. The first baseline network (denoted as "M1") is constructed by removing the temporal decoupling experts module and the teacher model, which means that only the supervised loss on labeled data is used for training. Then, we use the temporal decoupling experts module to replace the FFN module in the transformer block to build "M2". After that, "M3" is constructed based on "M2" by combining semi-supervised training with the unsupervised loss in Sect. 3.3.

Quantitative Comparison. Table 2 reports the quantitative scores of our method and the three baseline networks (*i.e.*, "M1" to "M3"). Specifically, compared with "M1", "M2" improves the PSNR score from 24.41 dB to 25.16 dB, the SSIM score from 0.9116 to 0.9217, and the LPIPS score from 0.0932 to 0.0822. This demonstrates the effectiveness of the temporal decoupling experts module in decomposing the physical components of snow videos in a temporal spirit, resulting in the enhanced recovery of background. Furthermore, our advanced "M3" model performs superior results compared to "M2", effectively showcasing the benefits of incorporating unlabeled data during the semi-supervised training to enhance the model's snow removal capabilities on synthetic and real data. Moreover, our network outperforms "M2" and "M3" in terms of the six metrics, which means leveraging the three components together enables the proposed network to achieve the best performance in video snow removal on both synthetic datasets and real-world datasets.

5 Conclusion

In this paper, we proposed the first semi-supervised video desnowing framework named SemiVDN, which effectively leverages knowledge from unlabeled data to enhance the generalization capabilities of deep models. To achieve superior snow removal, we incorporate the Prior-guided Temporal Decoupling Experts module, which explicitly decomposes the physical components of a snow video in a temporal manner. Furthermore, we propose a Distribution-driven Contrastive Regularization Loss that addresses the appearance discrepancy between synthetic and real data, ensuring the preservation of snow-invariant information. Observed from extensive experimentation on both synthesized and real-world snowy videos, our network demonstrates promising performance, surpassing existing state-of-the-art methods in snow removal.

Acknowledgements

This work is supported by the Guangzhou-HKUST(GZ) Joint Funding Program (No. 2023A03J0671), the National Natural Science Foundation of China (Grant No. 61902275), the Guangzhou Industrial Information and Intelligent Key Laboratory Project (No. 2024A03J0628), and Guangzhou-HKUST(GZ) Joint Funding Program (No. 2024A03J0618). This work is also supported by HKUST(GZ) College of Future Technology Red Bird MPhil Program and Yongjiang Technology (Ningbo) Co., Ltd.

References

1. Abbas, A., Andreopoulos, Y.: Biased mixtures of experts: Enabling computer vision inference under data transfer limitations. IEEE Transactions on Image Processing **29**, 7656–7667 (2020)
2. Ahmed, K., Baig, M.H., Torresani, L.: Network of experts for large-scale image categorization. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14. pp. 516–532. Springer (2016)
3. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. Advances in neural information processing systems **32** (2019)
4. Bossu, J., Hautiere, N., Tarel, J.P.: Rain or snow detection in image sequences through use of a histogram of orientation of streaks. International journal of computer vision **93**, 348–367 (2011)
5. Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C.: Basicvsr: The search for essential components in video super-resolution and beyond. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4947–4956 (2021)
6. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5972–5981 (2022)
7. Charbonnier, P., Blanc-Feraud, L., Aubert, G., Barlaud, M.: Two deterministic half-quadratic regularization algorithms for computed imaging. In: Proceedings of 1st international conference on image processing. vol. 2, pp. 168–172. IEEE (1994)
8. Chen, H., Ren, J., Gu, J., Wu, H., Lu, X., Cai, H., Zhu, L.: Snow removal in video: A new dataset and a novel method. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13211–13222 (2023)
9. Chen, S., Ye, T., Liu, Y., Bai, J., Chen, H., Lin, Y., Shi, J., Chen, E.: Cplformer: Cross-scale prototype learning transformer for image snow removal. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 4228–4239 (2023)
10. Chen, S., Ye, T., Liu, Y., Chen, E., Shi, J., Zhou, J.: Snowformer: Scale-aware transformer via context interaction for single image desnowing. arXiv preprint arXiv:2208.09703 (2022)
11. Chen, S., Ye, T., Xue, C., Chen, H., Liu, Y., Chen, E., Zhu, L.: Uncertainty-driven dynamic degradation perceiving and background modeling for efficient single image desnowing. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 4269–4280 (2023)

12. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
13. Chen, W.T., Fang, H.Y., Ding, J.J., Tsai, C.C., Kuo, S.Y.: Jstasr: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16. pp. 754–770. Springer (2020)
14. Chen, W.T., Fang, H.Y., Hsieh, C.L., Tsai, C.C., Chen, I., Ding, J.J., Kuo, S.Y., et al.: All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4196–4205 (2021)
15. Chen, X., Li, H., Li, M., Pan, J.: Learning a sparse transformer network for effective image deraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5896–5905 (2023)
16. Chen, Z., Zhu, L., Wan, L., Wang, S., Feng, W., Heng, P.A.: A multi-task mean teacher for semi-supervised shadow detection. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 5611–5620 (2020)
17. Fan, J., Weng, J., Wang, K., Yang, Y., Qian, J., Li, J., Yang, J.: Driving-video dehazing with non-aligned regularization for safety assistance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26109–26119 (2024)
18. Fan, Z., Sarkar, R., Jiang, Z., Chen, T., Zou, K., Cheng, Y., Hao, C., Wang, Z., et al.: M^3vit: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design. Advances in Neural Information Processing Systems **35**, 28441–28457 (2022)
19. Fang, Y., Wang, Z., Zhang, L., Cao, J., Chen, H., Xu, R.: Spiking wavelet transformer. arXiv preprint arXiv:2403.11138 (2024)
20. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. The Journal of Machine Learning Research **23**(1), 5232–5270 (2022)
21. Gross, S., Ranzato, M., Szlam, A.: Hard mixtures of experts for large scale weakly supervised vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6865–6873 (2017)
22. Gu, J., Cai, H., Chen, H., Ye, X., Ren, J., Dong, C.: Pipal: a large-scale image quality assessment dataset for perceptual image restoration. arXiv preprint arXiv:2007.12142 (2020)
23. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
24. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE transactions on pattern analysis and machine intelligence **33**(12), 2341–2353 (2010)
25. Huang, S., Wang, K., Liu, H., Chen, J., Li, Y.: Contrastive semi-supervised learning for underwater image restoration via reliable bank. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18145–18155 (2023)
26. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural computation **3**(1), 79–87 (1991)

27. Jiang, L., Dai, B., Wu, W., Loy, C.C.: Focal frequency loss for image reconstruction and synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13919–13929 (2021)
28. Jinjin, G., Haoming, C., Haoyu, C., Xiaoxing, Y., Ren, J.S., Chao, D.: Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. pp. 633–651. Springer (2020)
29. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 694–711. Springer (2016)
30. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5148–5157 (2021)
31. Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., Chen, Z.: Gshard: Scaling giant models with conditional computation and automatic sharding. arXiv preprint arXiv:2006.16668 (2020)
32. Li, L., Dong, Y., Ren, W., Pan, J., Gao, C., Sang, N., Yang, M.H.: Semi-supervised image dehazing. IEEE Transactions on Image Processing **29**, 2766–2779 (2019)
33. Li, M., Cao, X., Zhao, Q., Zhang, L., Meng, D.: Online rain/snow removal from surveillance videos. IEEE Transactions on Image Processing **30**, 2029–2044 (2021)
34. Liu, R., Deng, H., Huang, Y., Shi, X., Lu, L., Sun, W., Wang, X., Dai, J., Li, H.: Fuseformer: Fusing fine-grained information in transformers for video inpainting. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14040–14049 (2021)
35. Liu, Y., Zhu, L., Pei, S., Fu, H., Qin, J., Zhang, Q., Wan, L., Feng, W.: From synthetic to real: Image dehazing collaborating with unlabeled real data. In: Proceedings of the 29th ACM international conference on multimedia. pp. 50–58 (2021)
36. Liu, Y.F., Jaw, D.W., Huang, S.C., Hwang, J.N.: Desnownet: Context-aware deep network for snow removal. IEEE Transactions on Image Processing **27**(6), 3064–3073 (2018)
37. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
38. Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B.: Image restoration with mean-reverting stochastic differential equations. International Conference on Machine Learning (2023)
39. McCartney, E.J.: Optics of the atmosphere: scattering by molecules and particles. New York (1976)
40. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
41. Özdenizci, O., Legenstein, R.: Restoring vision in adverse weather conditions with patch-based denoising diffusion models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
42. Pei, S.C., Tsai, Y.T., Lee, C.Y.: Removing rain and snow in a single image using saturation and visibility features. In: 2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). pp. 1–6. IEEE (2014)
43. Puigcerver, J., Riquelme, C., Mustafa, B., Houlsby, N.: From sparse to soft mixtures of experts. arXiv preprint arXiv:2308.00951 (2023)

44. Ren, H., Zhou, Y., Zhu, J., Fu, H., Huang, Y., Lin, X., Fang, Y., Ma, F., Yu, H., Cheng, B.: Rethinking efficient and effective point-based networks for event camera classification and regression: Eventmamba. arXiv preprint arXiv:2405.06116 (2024)
45. Ren, W., Tian, J., Han, Z., Chan, A., Tang, Y.: Video desnowing and deraining based on matrix decomposition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4210–4219 (2017)
46. Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Susano Pinto, A., Keysers, D., Houlsby, N.: Scaling vision with sparse mixture of experts. Advances in Neural Information Processing Systems **34**, 8583–8595 (2021)
47. Shao, Y., Li, L., Ren, W., Gao, C., Sang, N.: Domain adaptation for image dehazing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2808–2817 (2020)
48. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017)
49. Shen, S., Hou, L., Zhou, Y., Du, N., Longpre, S., Wei, J., Chung, H.W., Zoph, B., Fedus, W., Chen, X., et al.: Flan-moe: Scaling instruction-finetuned language models with sparse mixture of experts. arXiv preprint arXiv:2305.14705 (2023)
50. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
51. Talebi, H., Milanfar, P.: Nima: Neural image assessment. IEEE transactions on image processing **27**(8), 3998–4011 (2018)
52. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
53. Valanarasu, J.M.J., Yasarla, R., Patel, V.M.: Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2353–2363 (2022)
54. Wang, H., Chen, J., Zhang, S., He, Y., Xu, J., Wu, M., He, J., Liao, W., Luo, X.: Dual-reference source-free active domain adaptation for nasopharyngeal carcinoma tumor segmentation across multiple hospitals. IEEE Transactions on Medical Imaging (2024)
55. Wang, H., Jin, Y., Zhu, L.: Dynamic interactive relation capturing via scene graph learning for robotic surgical report generation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 2702–2709. IEEE (2023)
56. Wang, H., Luo, X., Chen, W., Tang, Q., Xin, M., Wang, Q., Zhu, L.: Advancing uwf-slo vessel segmentation with source-free active domain adaptation and a novel multi-center dataset. arXiv preprint arXiv:2406.13645 (2024)
57. Wang, H., Zhang, S., Luo, X., Liao, W., Zhu, L.: Advancing delineation of gross tumor volume based on magnetic resonance imaging by performing source-free domain adaptation in nasopharyngeal carcinoma. In: International Workshop on Computational Mathematics Modeling in Cancer Analysis. pp. 71–80. Springer (2023)
58. Wang, H., Zhu, L., Yang, G., Guo, Y., Zhang, S., Xu, B., Jin, Y.: Video-instrument synergistic network for referring video instrument segmentation in robotic surgery. arXiv preprint arXiv:2308.09475 (2023)
59. Wang, L., Yoon, K.J.: Semi-supervised student-teacher learning for single image super-resolution. Pattern Recognition **121**, 108206 (2022)

60. Wang, X., Yu, F., Dunlap, L., Ma, Y.A., Wang, R., Mirhoseini, A., Darrell, T., Gonzalez, J.E.: Deep mixture of experts via shallow embedding. In: Uncertainty in artificial intelligence. pp. 552–562. PMLR (2020)
61. Wang, Y., Lin, S., Qu, Y., Wu, H., Zhang, Z., Xie, Y., Yao, A.: Towards compact single image super-resolution via contrastive self-distillation. arXiv preprint arXiv:2105.11683 (2021)
62. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
63. Wang, Z., Fang, Y., Cao, J., Zhang, Q., Wang, Z., Xu, R.: Masked spiking transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1761–1771 (2023)
64. Wei, W., Meng, D., Zhao, Q., Xu, Z., Wu, Y.: Semi-supervised transfer learning for image rain removal. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3877–3886 (2019)
65. Wu, H., Qu, Y., Lin, S., Zhou, J., Qiao, R., Zhang, Z., Xie, Y., Ma, L.: Contrastive learning for compact single image dehazing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10551–10560 (2021)
66. Wu, H., Yang, Y., Chen, H., Ren, J., Zhu, L.: Mask-guided progressive network for joint raindrop and rain streak removal in videos. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 7216–7225 (2023)
67. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. arXiv preprint arXiv:2401.13560 (2024)
68. Xing, Z., Zhu, L., Yu, L., Xing, Z., Wan, L.: Hybrid masked image modeling for 3d medical image segmentation. IEEE Journal of Biomedical and Health Informatics (2024)
69. Xu, J., Hu, X., Zhu, L., Dou, Q., Dai, J., Qiao, Y., Heng, P.A.: Video dehazing via a multi-range temporal alignment network with physical prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
70. Yang, Y., Aviles-Rivero, A.I., Fu, H., Liu, Y., Wang, W., Zhu, L.: Video adverse-weather-component suppression network via weather messenger and adversarial backpropagation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13200–13210 (2023)
71. Yang, Y., Fu, H., Aviles-Rivero, A.I., Schönlieb, C.B., Zhu, L.: Diffmic: Dual-guidance diffusion network for medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 95–105. Springer (2023)
72. Yang, Y., Wang, S., Liu, L., Hickman, S., Gilbert, F.J., Schönlieb, C.B., Aviles-Rivero, A.I.: Mammodg: Generalisable deep learning breaks the limits of cross-domain multi-center breast cancer screening. arXiv preprint arXiv:2308.01057 (2023)
73. Yang, Y., Wu, H., Aviles-Rivero, A.I., Zhang, Y., Qin, J., Zhu, L.: Genuine knowledge from practice: Diffusion test-time adaptation for video adverse weather removal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 25606–25616 (June 2024)
74. Yang, Y., Xing, Z., Zhu, L.: Vivim: a video vision mamba for medical video object segmentation. arXiv preprint arXiv:2401.14168 (2024)

75. Ye, Y., Chang, Y., Zhou, H., Yan, L.: Closing the loop: Joint rain generation and removal via disentangled image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2053–2062 (2021)
76. Yuan, K., Yu, Z., Liu, X., Xie, W., Yue, H., Yang, J.: Auformer: Vision transformers are parameter-efficient facial action unit detectors. arXiv preprint arXiv:2403.04697 (2024)
77. Yue, Z., Xie, J., Zhao, Q., Meng, D.: Semi-supervised video deraining with dynamical rain generator. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 642–652 (2021)
78. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5728–5739 (2022)
79. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14821–14831 (2021)
80. Zhang, K., Li, R., Yu, Y., Luo, W., Li, C.: Deep dense multi-scale network for snow removal using semantic and depth priors. IEEE Transactions on Image Processing **30**, 7419–7431 (2021)
81. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
82. Zheng, X., Liao, Y., Guo, W., Fu, X., Ding, X.: Single-image-based rain and snow removal using multi-guided filter. In: Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20. pp. 258–265. Springer (2013)
83. Zhu, X.J.: Semi-supervised learning literature survey (2005)